Project Title: SMS Spam Classification

Table of Contents

1. Introduction

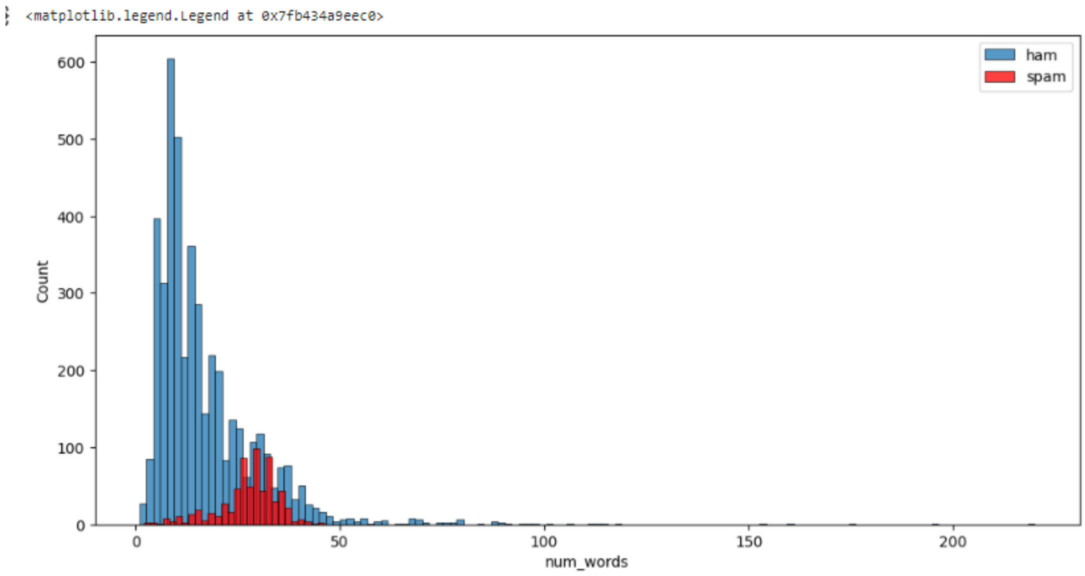
The SMS Spam Classification project focuses on building a machine learning model to classify SMS messages as either "spam" or "not spam" (ham). This report details the various stages of the project, from data cleaning and preprocessing to model building and deployment as a web application.

2. Data Cleaning

The initial dataset contained multiple unnecessary columns, including 'Unnamed: 2', 'Unnamed: 3', and 'Unnamed: 4', which were removed. Additionally, duplicate records were eliminated, resulting in a clean dataset of 5,169 SMS messages.

3. Exploratory Data Analysis (EDA)

- The dataset consists of 4,516 "ham" (not spam) and 653 "spam" messages, indicating class imbalance.
- Visualizations were created to explore the distribution of the number of characters, words, and sentences in the messages.



```
<matplotlib.legend.Legend at 0x7fb434a9eec0>
```

Figure 1: Histograms of Number of Words.

4. Data Preprocessing

Text preprocessing involved:
- Converting text to lowercase
- Tokenization
- Removing special characters, stopwords, and punctuation
- Stemming using the Porter Stemmer

5. Model Building

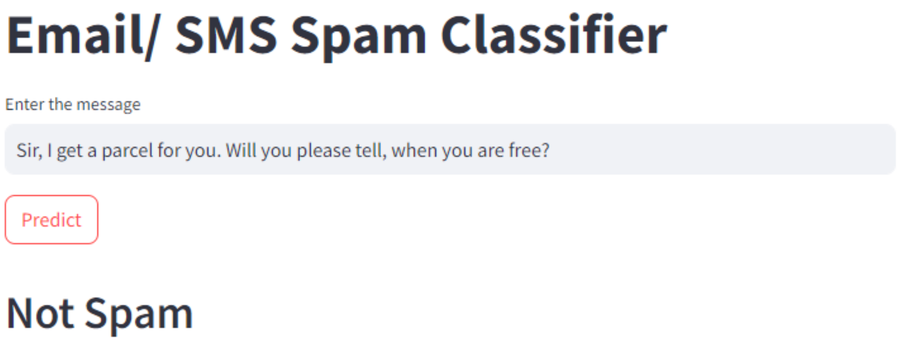
- Text vectorization was performed using TF-IDF (Term Frequency-Inverse Document Frequency).
- Various classification models were evaluated, including Multinomial Naive Bayes, Random Forest, AdaBoost, and others.
- The best-performing model, Multinomial Naive Bayes, achieved an accuracy of 97.20% and precision of 100% on the test dataset.

6. Model Evaluation and Improvement

- A voting classifier and stacking classifier were explored, resulting in further improvements in accuracy and precision.

7. Deployment as a Streamlit Web App

The final model was deployed as a web application using Streamlit. Users can input an SMS message, and the app predicts whether it is spam or not.



Email/ SMS Spam Classifier

Enter the message

Sir, I get a parcel for you. Will you please tell, when you are free?

Not Spam

Figure 2: Streamlit Web App

8. Conclusion and Future Work

In conclusion, the SMS Spam Classification project successfully developed a machine learning model to classify SMS messages as spam or not spam. The project also provided a user-friendly web application for real-time predictions.

Future improvements could include:

- Fine-tuning the model to handle class imbalance more effectively.
- Incorporating more advanced natural language processing techniques.
- Expanding the dataset to further enhance model performance.

This project demonstrates the potential of machine learning in addressing real-world problems and provides a practical application for SMS message classification.